US006029195A

United States Patent [19]

Herz

[11] Patent Number: 6,029,195

[45] Date of Patent: Feb. 22, 2000

[54] **SYSTEM FOR CUSTOMIZED ELECTRONIC IDENTIFICATION OF DESIRABLE OBJECTS**

[76] Inventor: **Frederick S. M. Herz**, Box 625 Canaan Valley, Davis, W. Va. 26260

[21] Appl. No.: **08/985,731**

[22] Filed: **Dec. 5, 1997**

Related U.S. Application Data

[63] Continuation-in-part of application No. 08/346,425, Nov. 29, 1994, Pat. No. 5,758,257.

[60] Provisional application No. 60/032,461, Dec. 9, 1996.

[51] **Int. Cl.**[7] **G06F 15/16**; H04H 1/02; H04N 7/14

[52] **U.S. Cl.** **709/219**; 348/1; 455/2; 707/10

[58] **Field of Search** 395/200.47–200.49; 348/1, 2, 6, 7, 8, 10; 455/3.1, 4.1, 4.2, 5.1, 6.1, 6.2; 704/104; 709/217–219, 203; 707/10; H04N 7/10, 7/14, 7/173

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,706,080	11/1987	Sincoskie	340/825.02
5,245,656	9/1993	Loeb et al.	380/23
5,301,109	4/1994	Landauer et al.	364/419.19
5,321,833	6/1994	Chang et al.	395/600
5,331,554	7/1994	Graham	364/419.07
5,331,556	7/1994	Black, Jr. et al.	364/419.08
5,717,923	2/1998	Dedrick	704/104 X
5,724,567	3/1998	Rose et al.	707/10
5,754,939	5/1998	Herz et al.	455/4.2

OTHER PUBLICATIONS

"Scatter/Gather: A Cluster–based Approach to Browsing Large Document Collections" by Cutting et al., 15th Ann Int'l Sigir '92, ACM 318–329.

"Evolving Agents For Personalized Information Filtering", Sheth et al., Proc. 9th IEEE Conference on A1 for Applications.

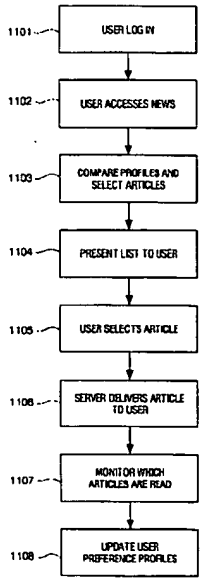
"A Secure And Privacy–Protecting Protocol For Transmitting Personal Information Between Organizations" Chaum et al.

Primary Examiner—John W. Miller
Attorney, Agent, or Firm—Duft, Graziano&Forest,P.C.

[57] **ABSTRACT**

This invention relates to customized electronic identification of desirable objects, such as news articles, in an electronic media environment, and in particular to a system that automatically constructs both a "target profile" for each target object in the electronic media based, for example, on the frequency with which each word appears in an article relative to its overall frequency of use in all articles, as well as a "target profile interest summary" for each user, which target profile interest summary describes the user's interest level in various types of target objects. The system then evaluates the target profiles against the users' target profile interest summaries to generate a user-customized rank ordered listing of target objects most likely to be of interest to each user so that the user can select from among these potentially relevant target objects, which were automatically selected by this system from the plethora of target objects that are profiled on the electronic media. Users' target profile interest summaries can be used to efficiently organize the distribution of information in a large scale system consisting of many users interconnected by means of a communication network. Additionally, a cryptographically-based pseudonym proxy server is provided to ensure the privacy of a user's target profile interest summary, by giving the user control over the ability of third parties to access this summary and to identify or contact the user.

15 Claims, 13 Drawing Sheets



number determined by calculating the statistical variance of the profiles of all target objects in a cluster, is termed a "cluster variance," (k.) a real number determined by calculating the maximum distance between the profiles of any two target objects in a cluster, is termed a "cluster diameter."

The system for electronic identification of desirable objects of the present invention automatically constructs both a target profile for each target object in the electronic media based, for example, on the frequency with which each word appears in an article relative to its overall frequency of use in all articles, as well as a "target profile interest summary" for each user, which target profile interest summary describes the user's interest level in various types of target objects. The system then evaluates the target profiles against the users' target profile interest summaries to generate a user-customized rank ordered listing of target objects most likely to be of interest to each user so that the user can select from among these potentially relevant target objects, which were automatically selected by this system from the plethora of target objects available on the electronic media.

Because people have multiple interests, a target profile interest summary for a single user must represent multiple areas of interest, for example, by consisting of a set of individual search profiles, each of which identifies one of the user's areas of interest. Each user is presented with those target objects whose profiles most closely match the user's interests as described by the user's target profile interest summary. Users' target profile interest summaries are automatically updated on a continuing basis to reflect each user's changing interests. In addition, target objects can be grouped into clusters based on their similarity to each other, for example, based on similarity of their topics in the case where the target objects are published articles, and menus automatically generated for each cluster of target objects to allow users to navigate throughout the clusters and manually locate target objects of interest. For reasons of confidentiality and privacy, a particular user may not wish to make public all of the interests recorded in the user's target profile interest summary, particularly when these interests are determined by the user's purchasing patterns. The user may desire that all or part of the target profile interest summary be kept confidential, such as information relating to the user's political, religious, financial or purchasing behavior; indeed, confidentiality with respect to purchasing behavior is the user's legal right in many states. It is therefore necessary that data in a user's target profile interest summary be protected from unwanted disclosure except with the user's agreement. At the same time, the user's target profile interest summaries must be accessible to the relevant servers that perform the matching of target objects to the users, if the benefit of this matching is desired by both providers and consumers of the target objects. The disclosed system provides a solution to the privacy problem by using a proxy server which acts as an intermediary between the information provider and the user. The proxy server dissociates the user's true identity from the pseudonym by the use of cryptographic techniques. The proxy server also permits users to control access to their target profile interest summaries and/or user profiles, including provision of this information to marketers and advertisers if they so desire, possibly in exchange for cash or other considerations. Marketers may purchase these profiles in order to target advertisements to particular users, or they may purchase partial user profiles, which do not include enough information to identify the individual users in question, in order to carry out standard kinds of demographic analysis and market research on the resulting database of partial user profiles. Pseudonymous control of an information server suggests how a special discount can be issued to a user's pseudonym and that such a digital credential is provided to the user as a result of his/her user profile making him/her eligible. The user may thus present this type of credential to the appropriate vendor to take advantage of the discount. This technique can be extended also to smart cards wherein the digital credential providing the discount is downloaded from the client to the smart card and upon presentation, the vendor may if desired, delete the credential upon redemption by the user. These discount credentials may similarly include any of the discount types (customized promotions) herein disclosed wherein each purchase may identified (characterized) and credentialized by the vendor onto the user's smart card and/or the vendor's system.

In the preferred embodiment of the invention, the system for customized electronic identification of desirable objects uses a fundamental methodology for accurately and efficiently matching users and target objects by automatically calculating, using and updating profile information that describes both the users' interests and the target objects' characteristics. The target objects may be published articles, purchasable items, or even other people, and their properties are stored, and/or represented and/or denoted on the electronic media as (digital) data. Examples of target objects can include, but are not limited to: a newspaper story of potential interest, a movie to watch, an item to buy, e-mail to receive, or another person to correspond with. In one suggested application, the user is a sender of email (which may have originated from the user for or from another external source such as from outside of a large organization) and the target objects are users who might be considered most appropriate based upon previous messages which they have received, read and responded to. Accordingly, like other target objects, users (or user pseudonyms) in accordance with their user profiles (or portions of which they have disclosed) may be organized and browsed within an automatically generated menu tree, which is below described in detail. In all these cases, the information delivery process in the preferred embodiment is based on determining the similarity between a profile for the target object and the profiles of target objects for which the user (or a similar user) has provided positive feedback in the past. The individual data that describe a target object and constitute the target object's profile are herein termed "attributes" of the target object. Attributes may include, but are not limited to, the following: (1) long pieces of text (a newspaper story, a movie review, a product description or an advertisement), (2) short pieces of text (name of a movie's director, name of town from which an advertisement was placed, name of the language in which an article was written), (3) numeric measurements (price of a product, rating given to a movie, reading level of a book), (4) associations with other types of objects (list of actors in a movie, list of persons who have read a document). Any of these attributes, but especially the numeric ones, may correlate with the quality of the target object, such as measures of its popularity (how often it is accessed) or of user satisfaction (number of complaints received).

The preferred embodiment of the system for customized electronic identification of desirable objects operates in an electronic media environment for accessing these target objects, which may be news, electronic mail, other published documents, or product descriptions. The system in its broadest construction comprises three conceptual modules, which may be separate entities distributed across many implementing systems, or combined into a lesser subset of physical entities. The specific embodiment of this system

disclosed herein illustrates the use of a first module which automatically constructs a "target profile" for each target object in the electronic media based on various descriptive attributes of the target object. A second module uses interest feedback from users to construct a "target profile interest summary" for each user, for example in the form of a "search profile set" consisting of a plurality of search profiles, each of which corresponds to a single topic of high interest for the user. The system further includes a profile processing module which estimates each user's interest in various target objects by reference to the users' target profile interest summaries, for example by comparing the target profiles of these target objects against the search profiles in users' search profile sets, and generates for each user a customized rank-ordered listing of target objects most likely to be of interest to that user. Each user's target profile interest summary is automatically updated on a continuing basis to reflect the user's changing interests.

Target objects may be of various sorts, and it is sometimes advantageous to use a single system that delivers and/or clusters target objects of several distinct sorts at once, in a unified framework. For example, users who exhibit a strong interest in certain novels may also show an interest in certain movies, presumably of a similar nature. A system in which some target objects are novels and other target objects are movies can discover such a correlation and exploit it in order to group particular novels with particular movies, e.g., for clustering purposes, or to recommend the movies to a user who has demonstrated interest in the novels. Similarly, if users who exhibit an interest in certain World Wide Web sites also exhibit an interest in certain products, the system can match the products with the sites and thereby recommend to the marketers of those products that they place advertisements at those sites, e.g., in the form of hypertext links to their own sites. The presently described system explains the techniques for target advertising (on a user by user basis) through both links from advertisements on a web page which tends to be visited by the most likely buyers of that particular product or service, and routing advertisements to such users via email. (This assumes that be cause user visitorship is measured at the level of the web page, certain pages within the web site may be more appropriate for certain advertisements due to the slight differences in its visitorship. Text chat(or acoustic voice chat) using a text to speech conversion module may be used in conjunction with real time profiling of the real time user dialogues occurring within that chat session. Advertisements which are relevant nature of the content being discussed at present may provide temporary links to the appropriate product such that when the nature of the content changes the advertisements changes (may disappear) accordingly.

The ability to measure the similarity of profiles describing target objects and a user's interests can be applied in two basic ways: filtering and browsing. Filtering is useful when large numbers of target objects are described in the electronic medias pace. These target objects can for example be articles that are received or potentially received by a user, who only has time to read a small fraction of them. For example, one might potentially receive all items on the AP news wire service, all items posted to a number of news groups, all advertisements in a set of newspapers, or all unsolicited electronic mail, but few people have the time or inclination to read so many articles. A filtering system in the system for customized electronic identification of desirable objects automatically selects a set of articles that the user is likely to wish to read. The accuracy of this filtering system improves over time by noting which articles the user reads and by generating a measurement of the depth to which the user reads each article. This information is then used to update the user's target profile interest summary. Browsing provides an alternate method of selecting a small subset of a large number of target objects, such as articles. Articles are organized so that users can actively navigate among groups of articles by moving from one group to a larger, more general group, to a smaller, more specific group, or to a closely related group. Each individual article forms a one-member group of its own, so that the user can navigate to and from individual article s as well as larger groups. The methods used by the system for customized electronic identification of desirable objects allow articles to be grouped into clusters and the clusters to be grouped and merged into larger and larger clusters. These hierarchies of clusters then form the basis for menuing and navigational systems to allow the rapid searching of large numbers of articles. This same clustering technique is applicable to any type of target objects that can be profiled on the electronic media such as product selections within a menu or throughout the World Wide Web.

There are a number of variations on the theme of developing and using profiles for article retrieval. Variations of this basic system are disclosed and comprise a system to filter electronic mail, an extension for retrieval of target objects such as purchasable items which may have more complex descriptions, a system to automatically build and alter menuing systems for browsing and searching through large numbers of target objects, and a system to construct virtual communities of people with common interests. These intelligent filters and browsers are necessary to provide a truly passive, intelligent system interface. A user interface that permits intuitive browsing and filtering represents for the first time an intelligent system for determining the affinities between users and target objects. The detailed, comprehensive target profiles and user-specific target profile interest summaries enable the system to provide responsive routing of specific queries for user information access. The information maps so produced and the application of users' target profile interest summaries to predict the information consumption patterns of a user allows for pre-caching of data at locations on the data communication network and at times that minimize the traffic flow in the communication network to thereby efficiently provide the desired information to the user and/or conserve valuable storage space by only storing those target objects (or segments thereof) which are relevant to the user's interests.

BRIEF DESCRIPTION OF THE DRAWING

FIG. **1** illustrates in block diagram form a typical architecture of an electronic media system in which the system for customized electronic identification of desirable objects of the present invention can be implemented as part of a user server system;

FIG. **2** illustrates in block diagram form one embodiment of the system for customized electronic identification of desirable objects;

FIGS. **3** and **4** illustrate typical network trees;

FIG. **5** illustrates in flow diagram form a method for automatically generating article profiles and an associated hierarchical menu system;

FIGS. **6–9** illustrate examples of menu generating process;

FIG. **10** illustrates in flow diagram form the operational steps taken by the system for customized electronic identification of desirable objects to screen articles for a user;

FIG. 11 illustrates a hierarchical cluster tree example;

FIG. 12 illustrates in flow diagram form the process for determination of likelihood of interest by a specific user in a selected target object;

FIGS. 13A–B illustrate in flow diagram form the automatic clustering process;

FIG. 14 illustrates in flow diagram form the use of the pseudonymous server;

FIG. 15 illustrates in flow diagram form the use of the system for accessing information in response to a user query; and

FIG. 16 illustrates in flow diagram form the use of the system for accessing information in response to a user query when the system is a distributed network implementation.

DETAILED DESCRIPTION

MEASURING SIMILARITY

This section describes a general procedure for automatically measuring the similarity between two target objects, or, more precisely, between target profiles that are automatically generated for each of the two target objects. This similarity determination process is applicable to target objects in a wide variety of contexts. Target objects being compared can be, as an example but not limited to: textual documents, human beings, movies, or mutual funds. It is assumed that the target profiles which describe the target objects are stored at one or more locations in a data communication network on data storage media associated with a computer system.

The computed similarity measurements serve as input to additional processes, which function to enable human users to locate desired target objects using a large computer system. These additional processes estimate a human user's interest in various target objects, or else cluster a plurality of target objects in to logically coherent groups. The methods used by these additional processes might in principle be implemented on either a single computer or on a computer network. Jointly or separately, they form the underpinning for various sorts of database systems and information retrieval systems.

Target Objects and Attributes

In classical Information Retrieval (IR) technology, the user is a literate human and the target objects in question are textual documents stored on data storage devices interconnected to the user via a computer network. That is, the target objects consist entirely of text, and so are digitally stored on the data storage devices within the computer network. However, there are other target object domains that present related retrieval problems that are not capable of being solved by present information retrieval technology which are applicable to targeting of articles and advertisements to readers of an on-line newspaper:

(a.) the user is a film buff and the target objects are movies available on videotape.

(b.) the user is a consumer and the target objects are used cars being sold.

(c.) the user is a consumer and the target objects are products being sold through promotional deals.

(d.) the user is an investor and the target objects are publicly traded stocks, mutual funds and/or real estate properties.

(e.) the user is a student and the target objects are classes being offered.

(f.) the user is an activist and the target objects are Congressional bills of potential concern.

(g.) the user is about to send an e-mail message and the target objects are potential recipients who are interested in the content of that message.

(h.) the user is a corporate receptionist receiving incoming e-mail, voice mail or live telephone calls and the target objects are the employees which are the most qualified to handle those incoming media.

(i.) the user is a net-surfer and the target objects are links to pages, servers, or newsgroups available on the World Wide Web which are linked from pages and articles in the on-line newspaper.

(j.) the user is a philanthropist and the target objects are charities.

(k.) the user is ill and the target objects are ads for medical specialists.

(l.) the user is an employee and the target objects are classifieds for potential employers.

(m.) the user is an employer and the target objects are classifieds for potential employees.

(n.) the user is a lonely heart and the target objects are classifieds for potential conversation partners.

(o.) the user is in search of an expert and the target objects are users, with known retrieval habits, of an document retrieval system.

(p.) the user is in need of insurance and the target objects are classifieds for insurance policy offers.

In all these cases, the user wishes to locate some small subset of the target objects—such as the target objects that the user most desires to rent, buy, investigate, meet, read, give mammograms to, insure, and so forth. The task is to help the user identify the most interesting target objects, where the user's interest in a target object is defined to be a numerical measurement of the user's relative desire to locate that object rather than others.

The generality of this problem motivates a general approach to solving the information retrieval problems noted above. It is assumed that many target objects are known to the system for customized electronic identification of desirable objects, and that specifically, the system stores (or has the ability to reconstruct) several pieces of information about each target object. These pieces of information are termed "attributes":

collectively, they are said to form a profile of the target object, or a "target profile." For example, where the system for customized electronic identification of desirable objects is activated to identify selections of interest in a particular category of on-line products for review or purchase by the user, it can be appreciated that there are certain unique sets of attributes which are pertinent to the particular product category of choice. For the application as part of a movie critic column (where the system identifies novel titles and reviews which are most interesting to the user) the system is likely be concerned with the values of attributes such as these:

(a.) title of movie,

(b.) name of director,

(c.) Motion Picture Association of America (MPAA) child-appropriateness rating (0=G, 1=PG, . . .),

(d.) date of release,

(e.) number of stars granted by a particular critic,

(f.) number of stars granted by a second critic,

(g.) number of stars granted by a third critic,

For example, a customized financial news column may be presented to the user in the form of articles which are of

interest to the user. In this case, however, an accordingly those stocks which are most interesting to the user may be presented as well.

(h.). full text of review by the third critic,

(i.). list of customers who have previously rented this movie,

(j.) list of actors.

Each movie has a different set of values for these attributes. This example conveniently illustrates three kinds of attributes. Attributes c–g are numeric attributes, of the sort that might be found in a database record. It is evident that they can be used to help the user identify target objects (movies) of interest. For example, the user might previously have rented many Parental Guidance (PG) films, and many films made in the 1970's. This generalization is useful: new films with values for one or both attributes that are numerically similar to these (such as MPAA rating of 1, release date of 1975) are judged similar to the films the user already likes, and therefore of probable interest. Attributes a–b and h are textual attributes. They too are important for helping the user locate desired films. For example, perhaps the user has shown a past interest in films whose review text (attribute h) contains words like "chase," "explosion," "explosions," "hero," "gripping," and "superb." This generalization is again useful in identifying new films of interest. Attribute i is an associative attribute. It records associations between the target objects in this domain, namely movies, and ancillary target objects of an entirely different sort, namely humans. A good indication that the user wants to rent a particular movie is that the user has previously rented other movies with similar attribute values, and this holds for attribute I just as it does for attributes a–h. For example, if the user has often liked movies that customer C_{17} and customer C_{190} have rented, then the user may like other such movies, which have similar values for attribute i. Attribute j is another example of an associative attribute, recording associations between target objects and actors. Notice that any of these attributes can be made subject to authentication when the profile is constructed, through the use of digital signatures; for example, the target object could be accompanied by a digitally signed note from the MPAA, which note names the target object and specifies its authentic value for attribute c.

These three kinds of attributes are common: numeric, textual, and associative. In the classical information retrieval problem, where the target objects are documents (or more generally, coherent document sections extracted by a text segmentation method), the system might only consider a single, textual attribute when measuring similarity: the full text of the target object. However, a more sophisticated system would consider a longer target profile, including numeric and associative attributes:

(a.) full text of document (textual),

(b.) title (textual),

(c.) author (textual),

(d.) language in which document is written (textual),

(e.) date of creation (numeric),

(f.) date of last update (numeric),

(g.) length in words (numeric),

(h.) reading level (numeric),

(i.) quality of document as rated by a third party editorial agency (numeric),

(j.) list of other readers who have retrieved this document (associative).

As another domain example, consider a domain where the user is an advertiser and the target objects are potential customers. The system might store the following attributes for each target object (potential customer):

(a.) first two digits of zip code (textual),

(b.) first three digits of zip code (textual),

(c.) entire five-digit zip code (textual),

(d.) distance of residence from advertiser's nearest physical storefront (numeric),

(e.) annual family income (numeric),

(f.) number of children (numeric),

(g.) list of previous items purchased by this potential customer (associative),

(h.) list of filenames stored on this potential customer's client computer (associative),

(i.) list of movies rented by this potential customer (associative),

(j.) list of investments in this potential customer's investment portfolio (associative),

(k.) list of documents retrieved by this potential customer (associative),

(l.) written response to Rorschach inkblot test (textual),

(m.) multiple-choice responses by this customer to 20 self-image questions (20 textual attributes).

As always, the notion is that similar consumers buy similar products. It should be noted that diverse sorts of information are being used here to characterize consumers, from their consumption patterns to their literary taste s and psychological peculiarities, and that this fact illustrates both the flexibility and power of the system for customized electronic identification of desirable objects of the present invention. Diverse sorts of information can be used as attributes in other domains as well (as when physical, economic, psychological and interest-related questions are used to profile the applicants to a dating service, which is indeed a possible domain for the present system), and the advertiser domain is simply an example.

As a final domain example, consider a domain where the user is an stock market investor and the target objects are publicly traded corporations. A great many attributes might be used to characterize each corporation, including but not limited to the following:

(a.) type of business (textual),

(b.) corporate mission statement (textual),

(c.) number of employees during each of the last 10 years (ten separate numeric attributes),

(d.) percentage growth in number of employees during each of the last 10 years,

(e.) dividend payment issued in each of the last 40 quarters, as a percentage of current share price,

(f.) percentage appreciation of stock value during each of the last 40 quarters, list of shareholders (associative),

(g.) composite text of recent articles about the corporation in the financial press (textual).

For example, a customized financial news column may be presented to the user in the form of articles which are of interest to the user. In addition, those stocks which are most interesting to the user may be presented as well.

It is worth noting some additional attributes that are of interest in some domains. In the case of documents and certain other domains, it is useful to know the source of each target object (for example, refereed journal article vs. UPI newswire article vs. Usenet newsgroup posting vs. question-answer pair from a question-and-answer list vs. tabloid newspaper article vs. . . .); the source may be represented

as a single-term textual attribute. Important associative attributes for a hypertext document are the list of documents that it links to, and the list of documents that link to it. Documents with similar citations are similar with respect to the former attribute, and documents that are cited in the same places are similar with respect to the latter. A convention may optionally be adopted that any document also links to itself. Especially in systems where users can choose whether or not to retrieve a target object, a target object's popularity (or circulation) can be usefully measured as a numeric attribute specifying the number of users who have retrieved that object. Related measurable numeric attributes that also indicate a kind of popularity include the number of replies to a target object, in the domain where target objects are messages posted to an electronic community such as an computer bulletin board or newsgroup, and the number of links leading to a target object, in the domain where target objects are interlinked hypertext documents on the World Wide Web or a similar system. A target object may also receive explicit numeric evaluations (another kind of numeric attribute) from various groups, such as the Motion Picture Association of America (MPAA), as above, which rates movies' appropriateness for children, or the American Medical Association, which might rate the accuracy and novelty of medical research papers, or a random survey sample of users (chosen from all users or a selected set of experts), who could be asked to rate nearly anything. Certain other types of evaluation, which also yield numeric attributes, may be carried out mechanically. For example, the difficulty of reading a text can be assessed by standard procedures that count word and sentence lengths, while the vulgarity of a text could be defined as (say) the number of vulgar words it contains, and the expertise of a text could be crudely assessed by counting the number of similar texts its author had previously retrieved and read using the invention, perhaps confining this count to texts that have high approval ratings from critics. Finally, it is possible to synthesize certain textual attributes mechanically, for example to reconstruct the script of a movie by applying speech recognition techniques to its soundtrack or by applying optical character recognition techniques to its closed-caption subtitles.

Decomposing Complex Attributes

Although textual and associative attributes are large and complex pieces of data, for information retrieval purposes they can be decomposed into smaller, simpler numeric attributes. This means that any set of attributes can be replaced by a (usually larger) set of numeric attributes, and hence that any profile can be represented as a vector of numbers denoting the values of these numeric attributes. In particular, a textual attribute, such as the full text of a movie review, can be replaced by a collection of numeric attributes that represent scores to denote the presence and significance of the words "aardvark," "aback," "abacus," and so on through "zymurgy" in that text. The score of a word in a text may be defined in numerous ways. The simplest definition is that the score is the rate of the word in the text, which is computed by computing the number of times the word occurs in the text, an d dividing this number by the total number of words in the text. This sort of score is often called the "term frequency" (TF) of the word. The definition of term frequency may optionally be modified to weight different portions of the text unequally: for example, any occurrence of a word in the text's title might be counted as a 3-fold or more generally k-fold occurrence (as if the title had been repeated k times within the text), in order to reflect a heuristic assumption that the words in the title are particularly important indicators of the text's content or topic.

However, for lengthy textual attributes, such as the text of an entire document, the score of a word is typically defined to be not merely its term frequency, but its term frequency multiplied by the negated logarithm of the word's "global frequency," as measured with respect to the textual attribute in question. The global frequency of a word, which effectively measures the word's uninformativeness, is a fraction between 0 and 1, defined to be the fraction of all target objects for which the textual attribute in question contains this word. This adjusted score is often known in the art as TF/IDF ("term frequency times inverse document frequency"). When global frequency of a word is taken into account in this way, the common, uninformative words have scores comparatively close to zero, no matter how often or rarely they appear in the text. Thus, their rate has little influence on the object's target profile. Alternative methods of calculating word scores include latent semantic indexing or probabilistic models.

Instead of breaking the text into its component words, one could alternatively break the text into overlapping word bigrams (sequences of 2 adjacent words), or more generally, word n-grams. These word n-grams may be scored in the same way as individual words. Another possibility is to use character n-grams. For example, this sentence contains a sequence of overlapping character 5-grams which starts "for e", "or ex", "r exa", "exam", "examp", etc. The sentence may be characterized, imprecisely but usefully, by the score of each possible character 5-gram ("aaaaa", "aaaab", . . . "zzzzz") in the sentence. Conceptually speaking, in the character 5-gram case, the textual attribute would be decomposed into at least 26^5=11,881,376 numeric attributes. Of course, for a given target object, most of these numeric attributes have values of 0, since most 5-grams do not appear in the target object attributes. These zero values need not be stored anywhere. For purposes of digital storage, the value of a textual attribute could be characterized by storing the set of character 5-grams that actually do appear in the text, together with the nonzero score of each one. Any 5-gram that is no t included in the set can be assumed to have a score of zero. The decomposition of textual attributes is not limited to attributes whose values are expected to be long texts. A simple, one-term textual attribute can be replaced by a collection of numeric attributes in exactly the same way. Consider again the case where the target objects are movies. The "name of director" attribute, which is textual, can be replaced by numeric attributes giving the scores for "Federico-Fellini," "Woody-Allen," "Terence-Davies," and so forth, in that attribute. For these one-term textual attributes, the score of a word is usually defined to be its rate in the text, without any consideration of global frequency. Note that under these conditions, one of the scores is 1, while the other scores are 0 and need not be stored. For example, if Davies did direct the film, then it is "Terence-Davies" whose score is 1, since "Terence-Davies" constitutes 100% of the words in the textual value of the "name of director" attribute. It might seem that nothing has been gained over simply regarding the textual attribute as having the string value "Terence-Davies." However, the trick of decomposing every non-numeric attribute into a collection of numeric attributes proves useful for the clustering and decision tree methods described later, which require the attribute values of different objects to be averaged and/or ordinally ranked. Only numeric attributes can be averaged or ranked in this way. Just as a textual attribute may be decomposed into a number of component terms (letter or word n-grams), an associative attribute may be decomposed into a number of component associations. For instance, in a